Measuring the Effectiveness of Facial Forensic Apprenticeships

P. Jonathon Phillips

Carina A. Hahn

National Institute of Standards and Technology

International Face Performance Conference

November 28, 2018

Outline

- Becoming a facial forensic examiner
  - What is involved
- Short-term training
  - What is known
- Facial forensics training
  - A proposed study

Not measured

Motivation for Proposed Study

- Efficacy of training: contentious
  - Psychology literature is on short term training and is overall negative
  - Facial forensic best practices recommends long term training

- Focus on accuracy

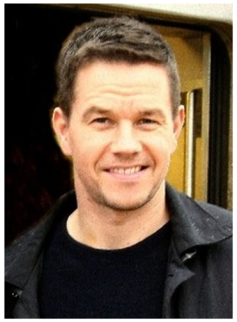
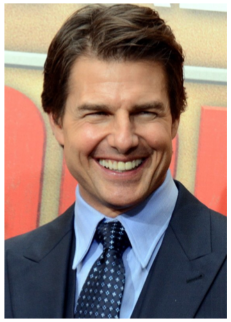
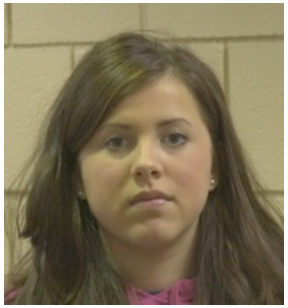
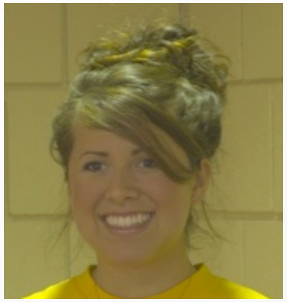
…there is more

What Do Facial Forensic Examiners Do?

- Compare two face images – determine whether same or different people
- Write detailed reports
- Testify in court
- Accurate and consistent
- Rigorous comparisons: hours to days

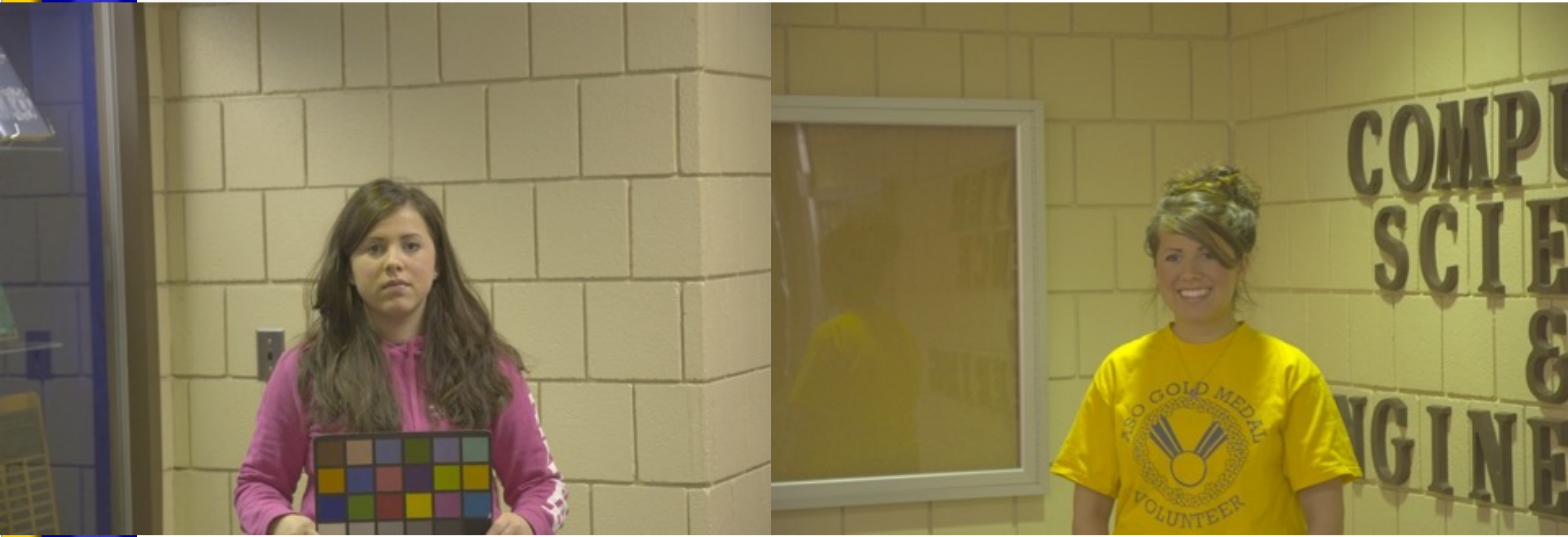
Wells (2012) Law 101: Legal Guide for the Forensic Expert. *NIJ Journal*

Face Recognition and Face Matching

	Familiar	Unfamiliar
Face Memory		
Face Matching	 	 

courtesy of Georges Biard & Metropolitan Transportation Authority of the State of New York & Peter Souza

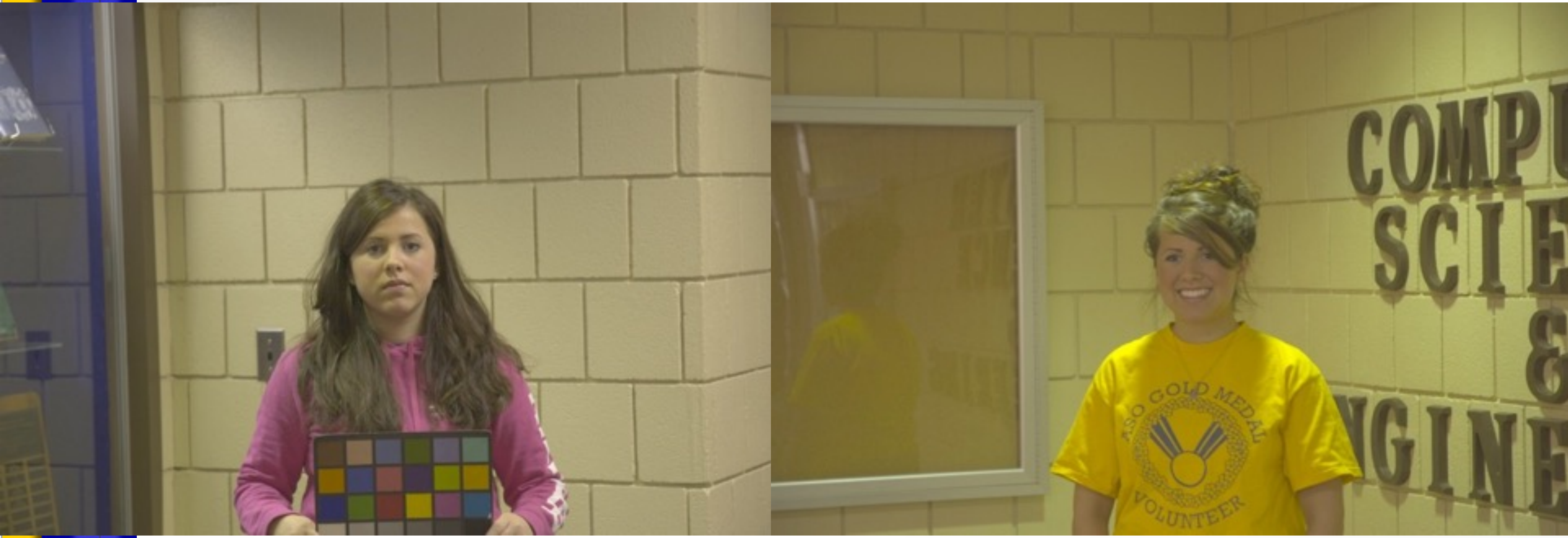
Face Matching



Same or Different?

Correct Answer:
Same

Face Matching



+3 The observations strongly support that it is the same person

+2 The observations support that it is the same person

+1 The observations support to some extent that it is the same person

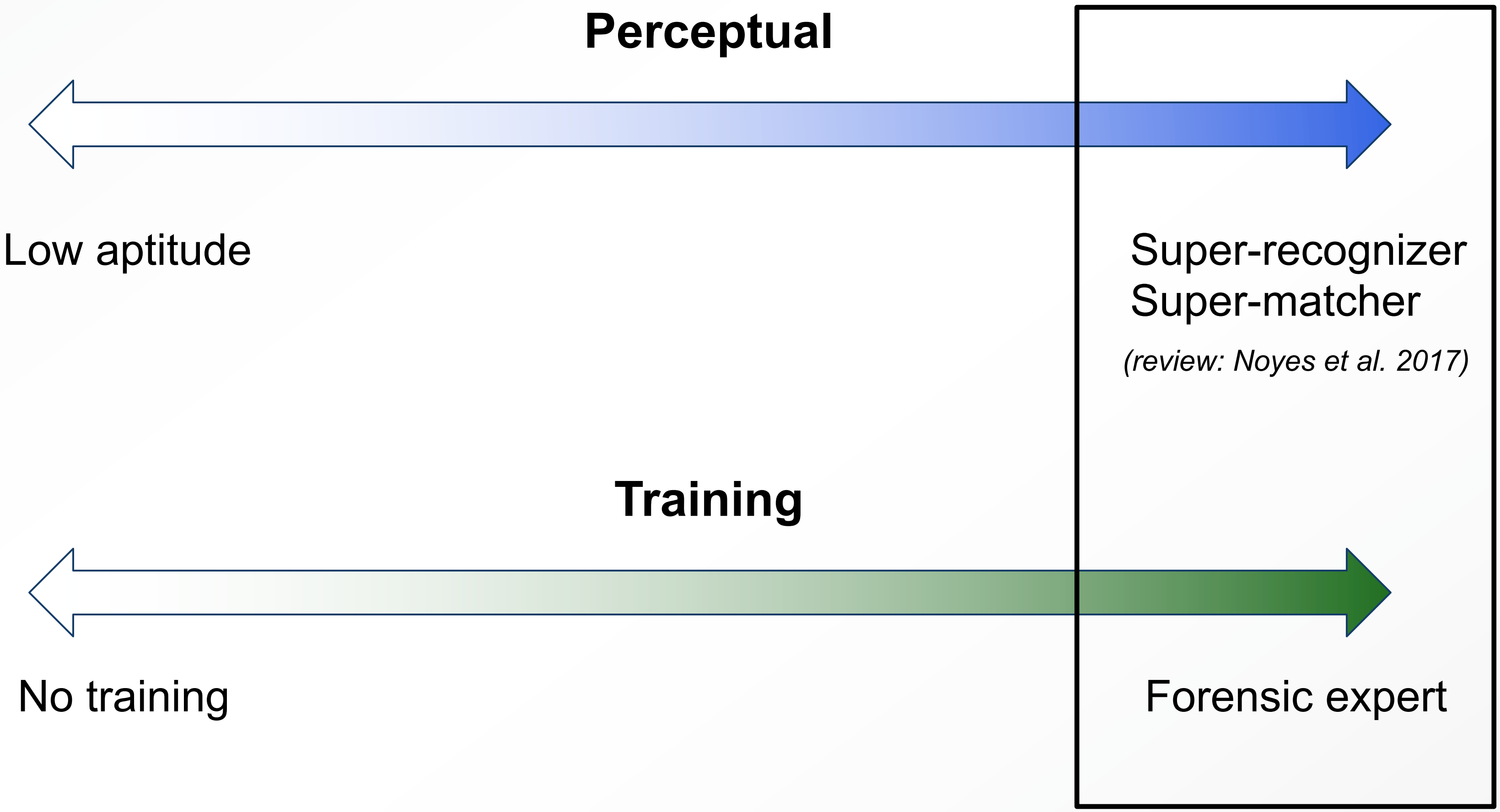
0 The observations support neither that it is the same person nor that it is different persons

-1 The observations support to some extent that it is not the same person

-2 The observations support that it is not the same person

-3 The observations strongly support that it is not the same person

Two Dimensions of Face Recognition & Matching



Perceptual

Low aptitude

Super-recognizer
Super-matcher

(review: Noyes et al. 2017)

Training

No training

Forensic expert

Are these the same?
What is independent benefit of training?

How to Become a Facial Forensic Examiner

- 1 – 4 year apprenticeship
- Intensive courses
- Mentoring

Facial Identification Working Group (2012) Recommendations for a Training Program in Facial Comparison

Goals of Apprenticeship

- Improve accuracy
- Improve consistency
  - Within person: same accuracy/judgments on different tests & days
  - Between people: rating scale consistency
- Learn to write reports and give testimony

Goals of Apprenticeship

- **Improve accuracy**
- Improve consistency
- Learn to write reports and give testimony

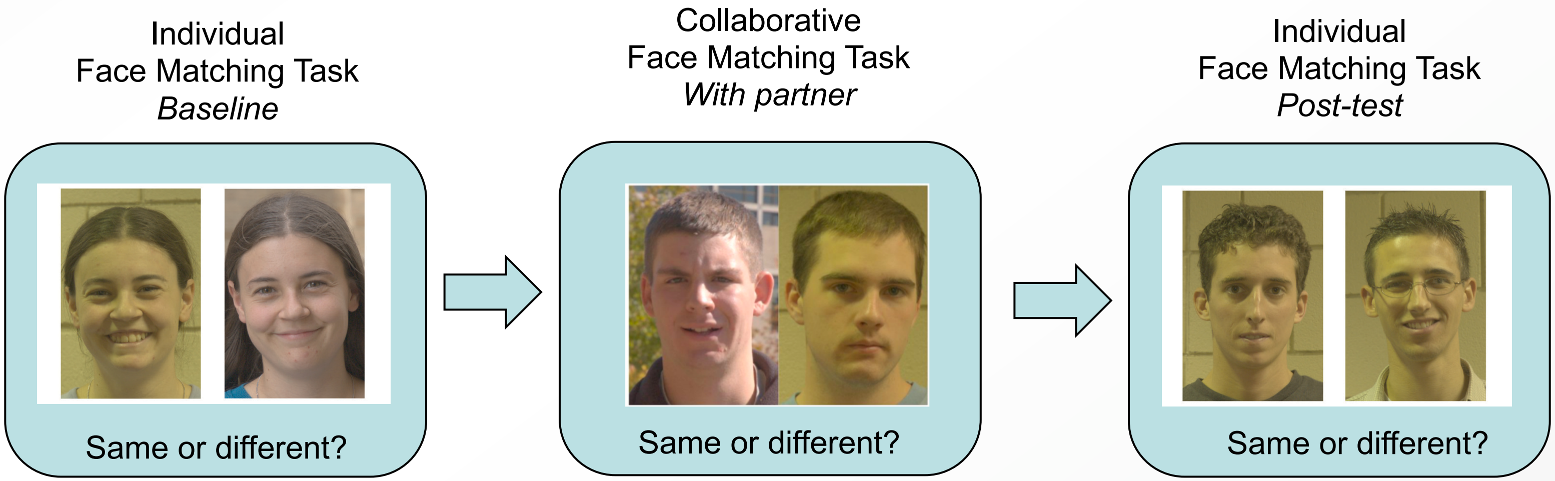
Methods that Improve Accuracy

- Accuracy
  - In-lab training that increases accuracy
    - Mentorship (Dowsett & Burton, 2015)
    - Feedback (White et al. 2014)
    - Feature comparison strategy (Megreya & Bindemann, 2018; Towler et al., 2017)

Mentorship

- Paradigm (Dowsett & Burton, 2015)

Baseline to Post-test:
Accuracy improved for low performers



Individual
Face Matching Task
Baseline

Same or different?

Collaborative
Face Matching Task
With partner

Same or different?

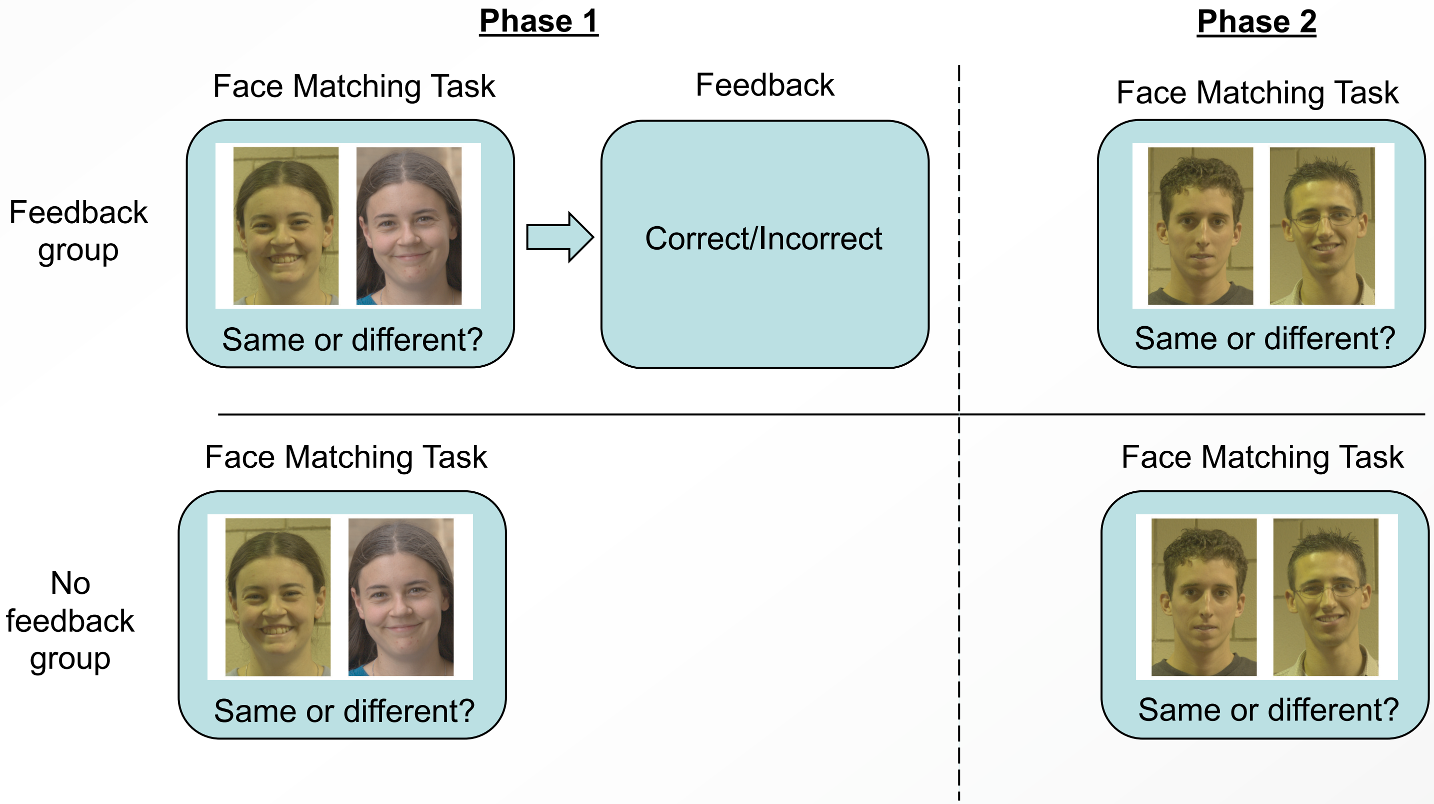
Individual
Face Matching Task
Post-test

Same or different?

Feedback

- Paradigm (White et al., 2014)

Phase 2:
Accuracy improved for low performers after feedback



Phase 1

Phase 2

Feedback group

Face Matching Task

Same or different?

Feedback

Correct/Incorrect

Face Matching Task

Same or different?

No feedback group

Face Matching Task

Same or different?

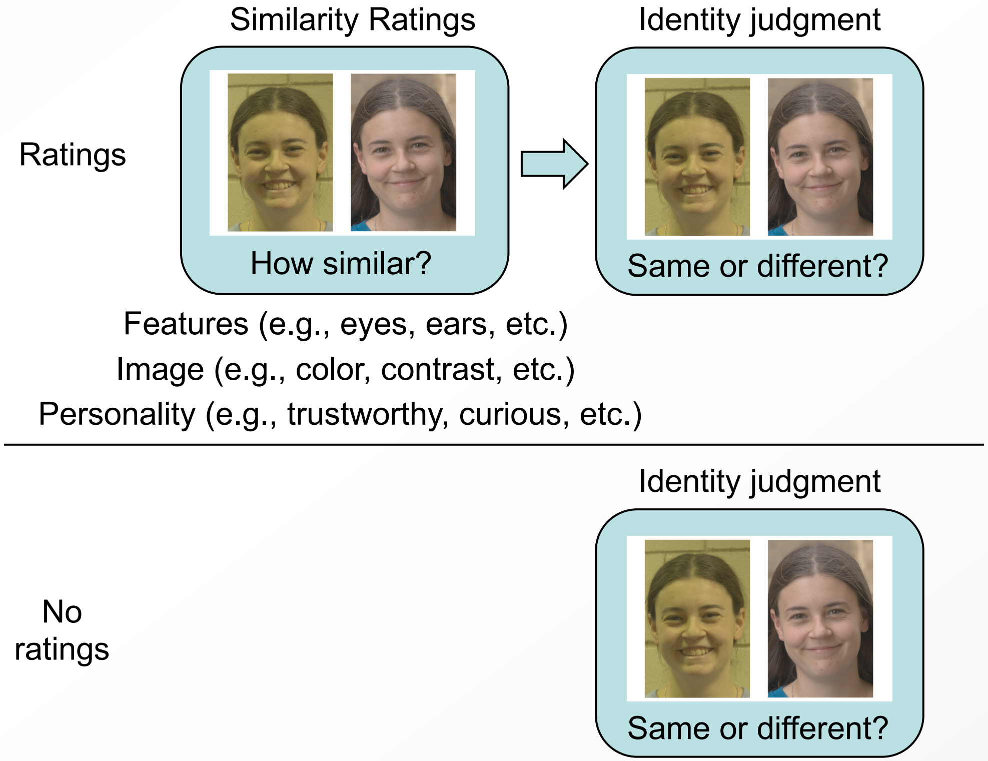
Face Matching Task

Same or different?

Feature Comparison Strategy

- Paradigm (Towler et al., 2017)

Rating feature or image similarity improved matching accuracy



Similarity Ratings

Identity judgment

Ratings

How similar?

Same or different?

Features (e.g., eyes, ears, etc.)
Image (e.g., color, contrast, etc.)
Personality (e.g., trustworthy, curious, etc.)

Identity judgment

No ratings

Same or different?

What is Known: Accuracy

- Accuracy
  - In-lab training that increases accuracy
    - Mentors (Dowsett & Burton, 2015)
    - Feedback (White et al. 2014)
    - Feature comparison strategy (Towler et al., 2017)
  - Caveats
    - All short-term training
      - Longest: face memory (29 days; Dolzycka et al., 2014)
    - Mentors & feedback: only lower performers benefit
    - Feature comparison strategy: Criterion shifts
- Long-term training: no studies

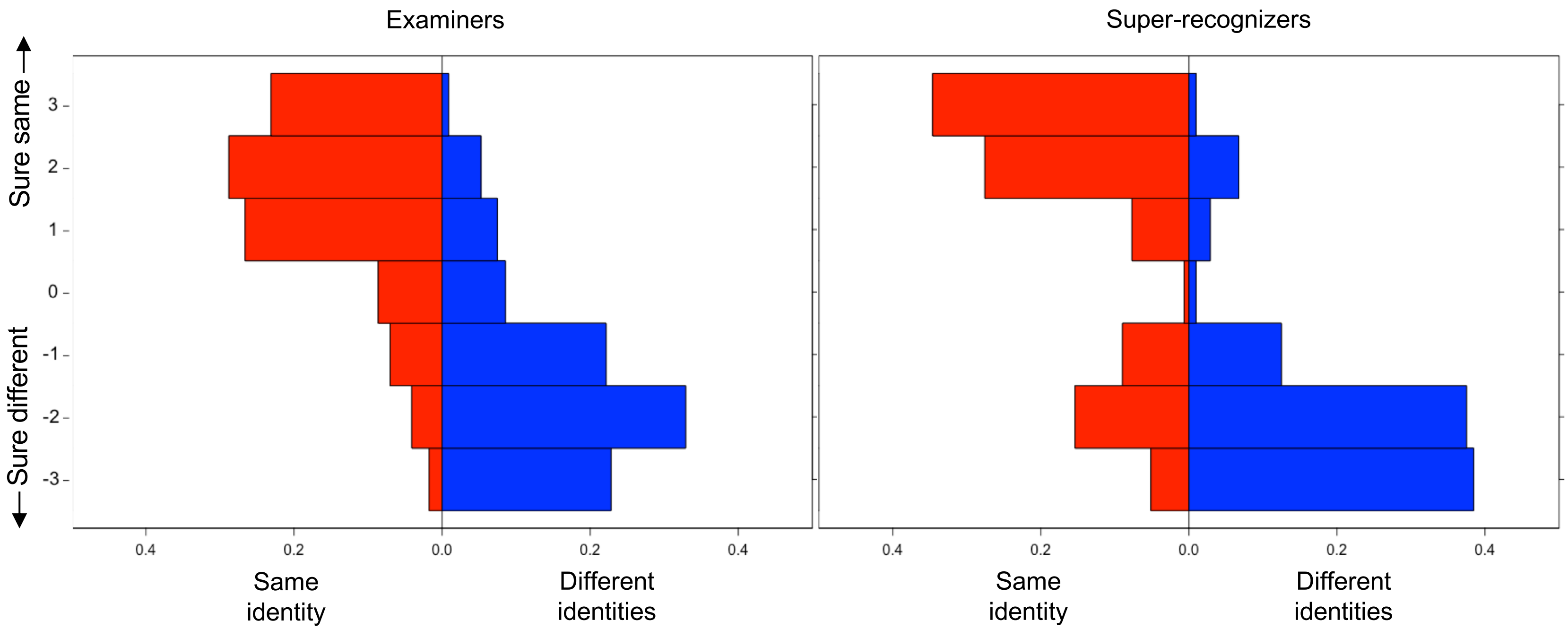
Goals of Apprenticeship

- Improve accuracy
- **Improve consistency**
- Learn to write reports and give testimony

Examiners vs. Super-recognizers

- Phillips et al., 2018
  - Both groups: higher face matching accuracy than untrained students
  - Examiners = Super-recognizers
- Comparison of examiners to super-recognizers
  - tease apart natural ability vs. training

Consistent Use of Rating Scale



Equal accuracy overall

Training may influence the way response scale is used

Re-analysis of data from Phillips et al. (2018)

Consistent Use of Rating Scale

- Within group consistency
  - – Inter-rater reliability (Fleiss’s Weighted Kappa)
    - Measure of agreement/consistency across participants

Participant

Face pair	1	2
A	5	5
B	4	4
C	1	1
D	3	3
	2	2
E	5	5

⬆ High Kappa Coefficient

Participant

Face pair	1	2
A	3	5
B	1	4
C	5	1
D	3	1
	2	5
E	1	5

⬇ Low Kappa Coefficient

Inter-rater Reliability
Fleiss's Weighted Kappa

- **Examiners = 0.40**; 95% CI [0.31, 0.49], *p* < .001
- **Super-recognizers = 0.28**; 95% CI [0.17, 0.39], *p* < .001

- Higher agreement among examiners compared to super-recognizers

Re-analysis of data from Phillips et al. (2018)

Consistent Use of Rating Scale

- Phillips et al., 2018
  - Different use of rating scale by facial examiners and super-recognizers
- Norell et al., 2014
  - Professional face examiners: more likely to respond “I don’t know” with poor quality images compared to untrained students

Goals of Apprenticeship

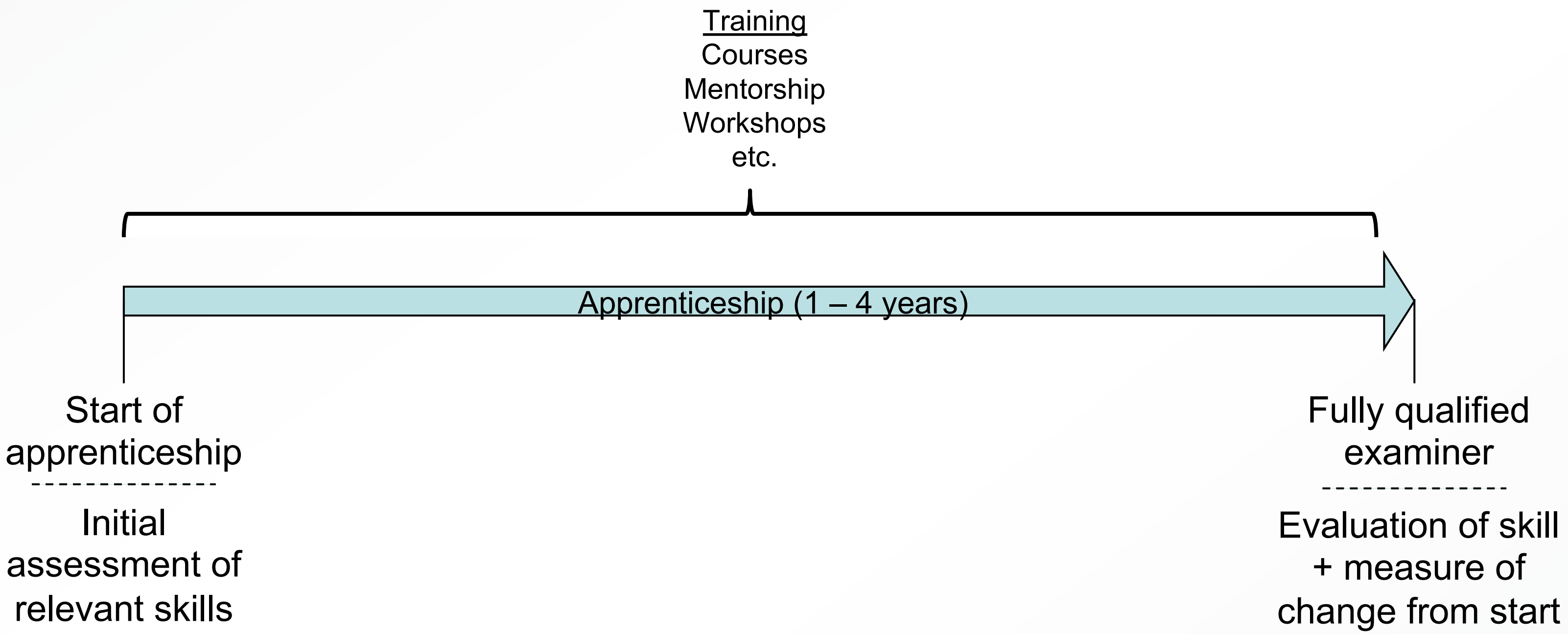
- Improve accuracy
- Improve consistency
- **Learn to write reports and give testimony**

Not measured

The Path Forward

- Proposed study: How to measure effects of training

How to Measure Effects of Training



Training
Courses
Mentorship
Workshops
etc.

Apprenticeship (1 – 4 years)

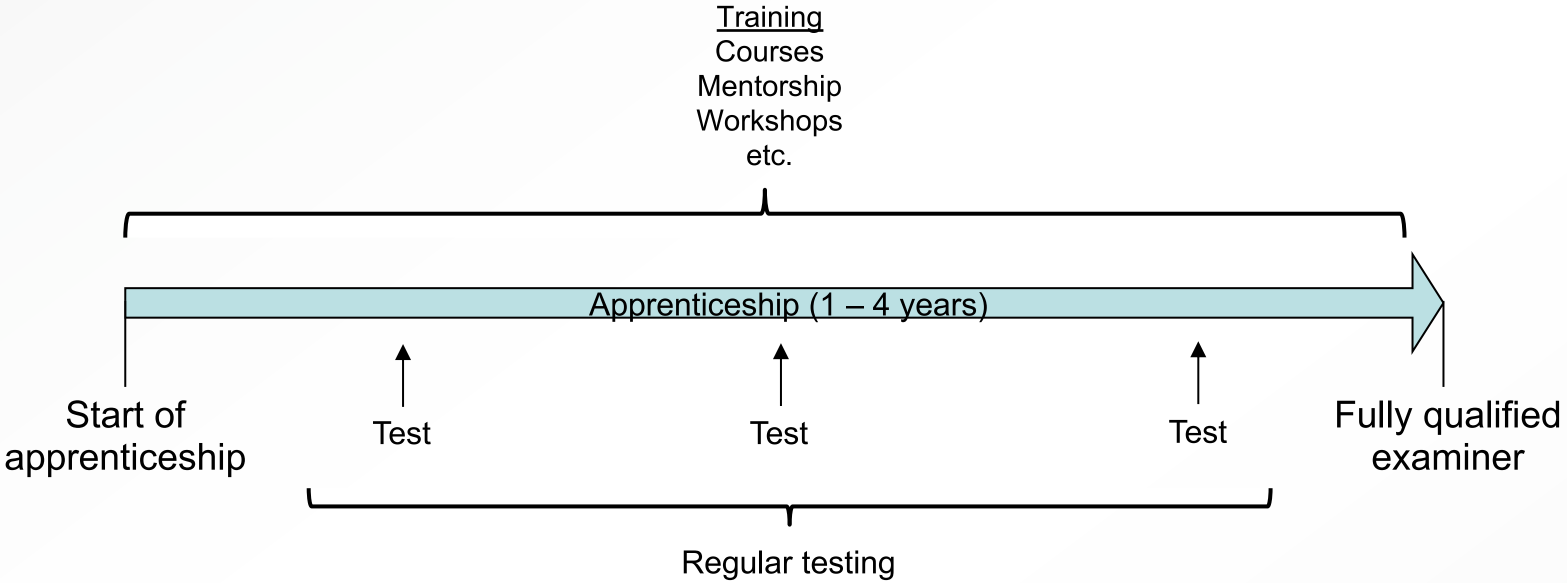
Start of apprenticeship

Initial assessment of relevant skills

Fully qualified examiner

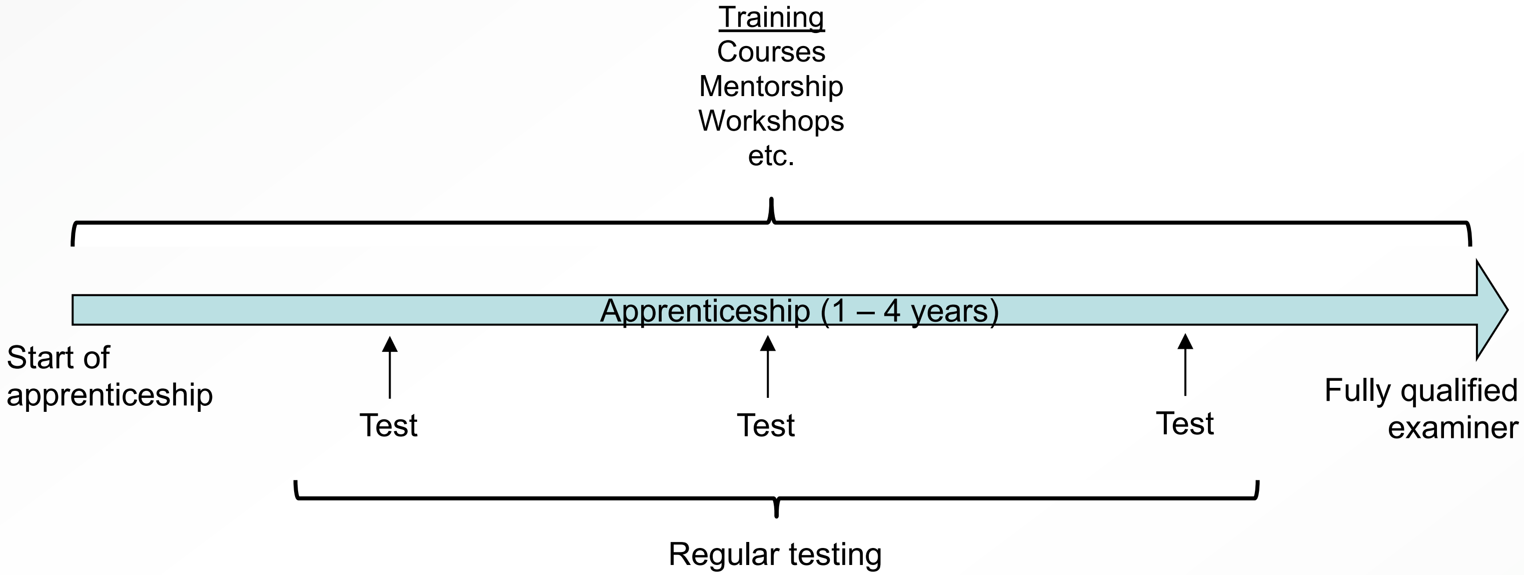
Evaluation of skill + measure of change from start

How to Measure Effects of Training



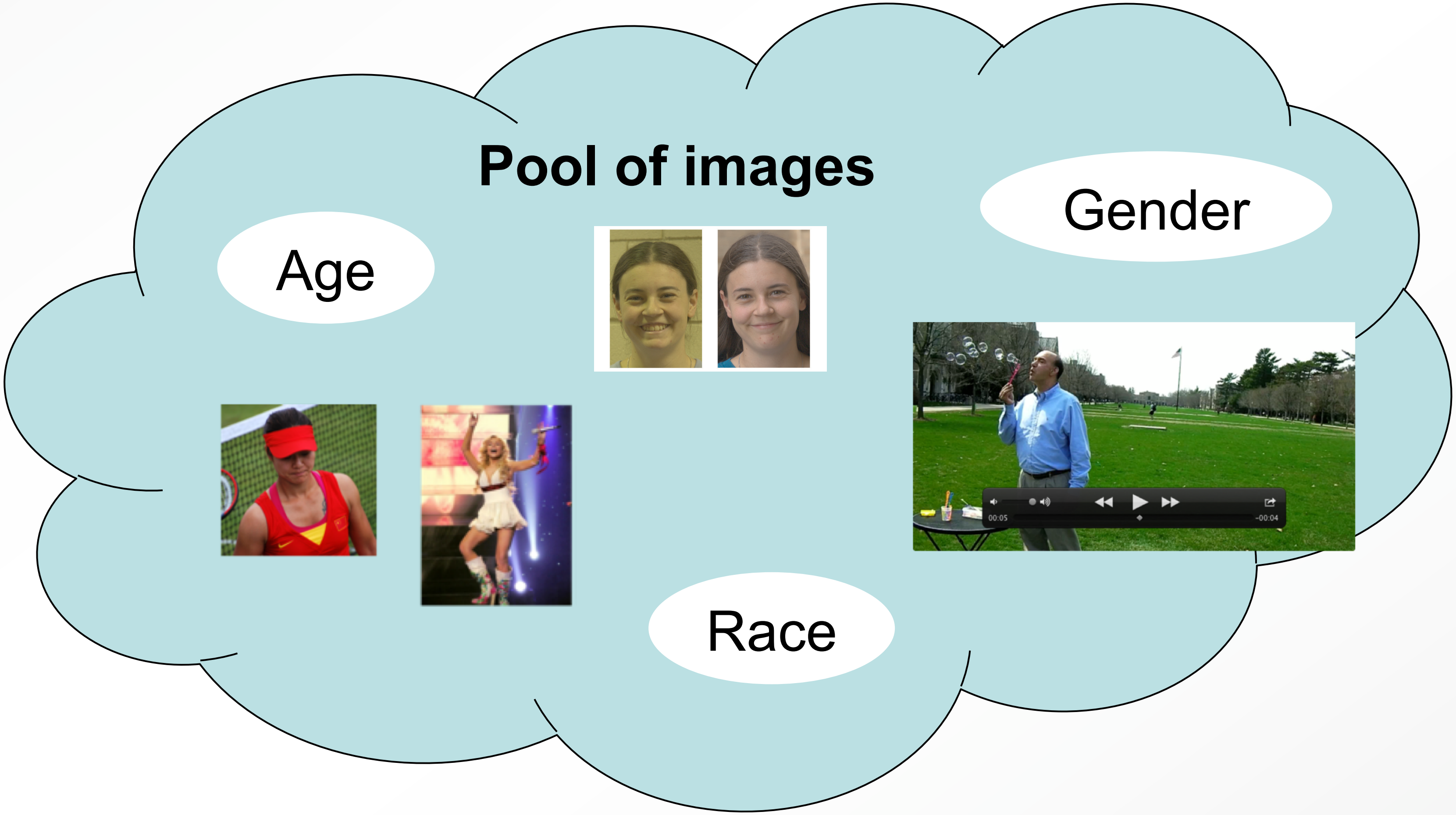
- Purpose of regular testing
  - Accuracy on relevant tasks
    - Change in performance over apprenticeship
  - Progress at regular intervals
    - Pinpoint key components of training

How to Measure Effects of Training



- Properties of tests
  - Measure change in skill: consistent difficulty throughout training
  - Tasks representative of forensic casework
  - Write reports
  - Outcome: metrics that quantify abilities
  - Multiple metrics are necessary

The Path Forward



Pool of images

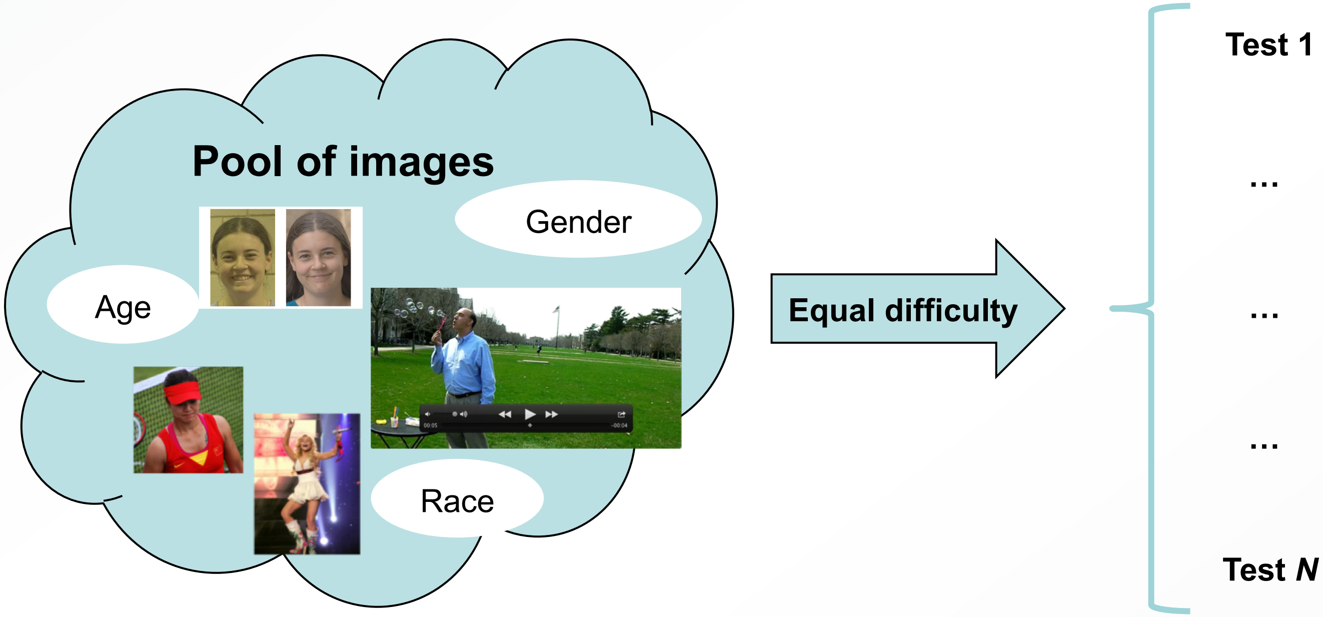
Age

Gender

Race

IARPA Janus Benchmark-B Face Dataset. C. Whitelam, et al CVPR, Workshop on Biometrics, 2017.

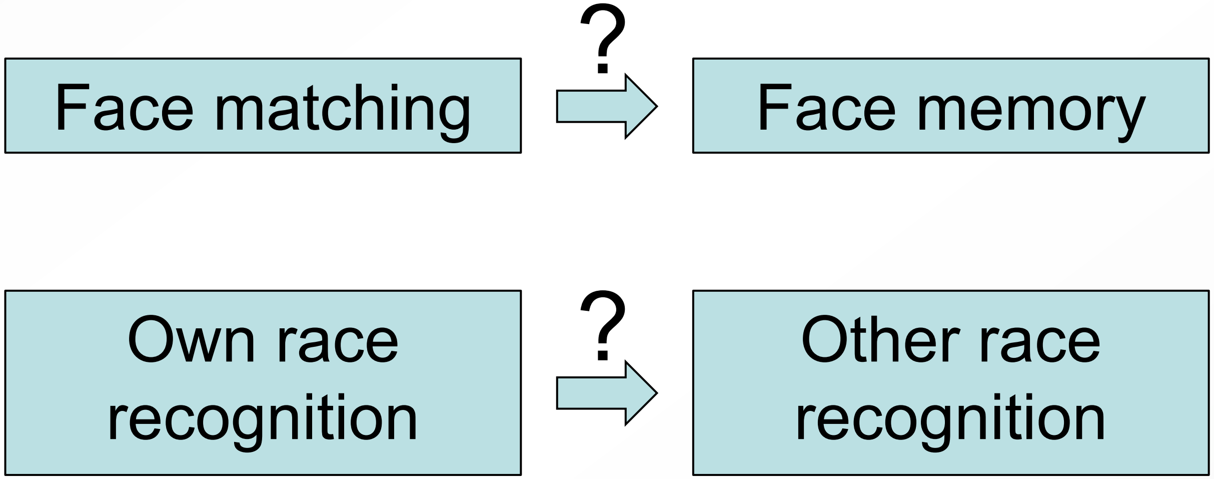
The Path Forward



- Large database
  - No repetition of images (familiarity)
  - Reflective of casework
  - Sufficient difficulty
  - Racial/ethnic diversity that reflects underlying population

The Path Forward

- Relationship between tests



Face matching ? Face memory

Own race recognition ? Other race recognition

Balsdon et al., 2018; Bate et al., 2018; Bate et al. 2018a; Noyes et al., 2018

Benefits to community

- Initial assessment
  - What level of ability acceptable?
- Testing at regular intervals
  - Assess critical elements of training
- Consistency
  - Across facial forensic community
- Increased ability of facial examiners

Summary

- Training: What is known to work
  - Mentorship (Dowsett & Burton, 2015)
  - Feedback (White et al., 2014)
  - Feature comparisons (Towler et al., 2017)
  - Short-term (< 1 month)
- No evaluations of long-term training
- Path forward
  - Battery of tests
    - calibrate to equal difficulty
    - compare across tasks
    - reflect casework
    - test at regular intervals
    - measure long-term

To be measured

Questions?